International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

Application of GA in Web Quality Estimation for Regression Model Optimization

Shama Bano
Department of CSE-SE
BBDU, Lucknow
shamabano00@gmail.com

Smriti Rastogi
Department of Computer Science & Engg.
BBDU, Lucknow
smritirastogibbd@gmail.com

***Abstract*-- The rapid development in the database technology, especially in the storage capacity has realized the databases with the size of terabytes, consequently the current databases may also have a huge number of the attributes. The aim of our work is to propose an improved method for mining and pruning association rules based on evolutionary computational methods such as Genetic Algorithms (GA) and Genetic Relation Algorithm (GRA). Furthermore, to demonstrate the usefulness of GA as a selection method, GA is applied to Web mining, where one of the major problems is the poor quality of the information retrieved because of the lack of performing additional computation.**

Keywords--Genetic Algorithm, Web Mining, Estimation analysis, Data prediction.

1 Introduction

Web mining is the automated extraction of predictive information from large databases. It is an interdisciplinary field involving databases, machine learning, statistics, artificial intelligence, information retrieval and visualization in order to extract high level knowledge from real-world data sets [1]. Data mining is the core step of the knowledge discovery process. It has attracted a lot of interest in the database community mainly motivated by the explosive growth and availability of huge amount of data in related to science, business, medicine, government, etc. Through data mining, it is possible to find novel and useful patterns from databases. Many data mining techniques have been developed since the last decade, that go beyond simple analysis, contributing greatly to business strategies, knowledge bases and scientific research. Furthermore, data mining has become a key topic in database technology.

Web Mining has emerged as an important research area and association rule mining becomes one of the most attractive tasks of data mining, nevertheless, there are still some unsolved problems. A problem of classical association rule mining algorithms is that they are mostly limited to deal with a relative small size of databases, in the number of records and in the number of attributes as well. However, the rapid development in the database technology, especially in the storage capacity has realized the databases with the size of terabytes, consequently the current databases may also have a huge number of the attributes. The aim of our work is to propose an improved method for mining and pruning association rules based on evolutionary computational methods such as Genetic Algorithms (GA) and Genetic Relation Algorithm (GRA). Furthermore, to demonstrate the usefulness of GA as a selection method, GA is applied to Web mining, where one of the major problems is the poor quality of the information retrieved because of the lack of performing additional computation.

Association Rule mining has received great interest by the scientific community since it is very convenient to find patterns or rules on which attributes that occur frequently together in a given dataset. Association rule mining has been introduced in 1993 by Agrawal et al. [2] and it is continuously been applied to several fields such as medicine, finance, stock market, manufacturing, e-business, intrusion detection, bio-informatics, etc.

In the last decade, several techniques have been applied to association rule mining such as artificial neural networks (ANN), expert systems (ES), linear programming (LP), database systems (DS), and evolutionary computing (EC). EC is a computational technique inspired by the natural evolution process that imitates the mechanism of natural selection and survival of the fittest. When EC is applied to data mining, they provide vigorous search method which performs a global search in the candidate solution space. On the other hand, the majority of association rules perform a local search in the candidate space. Instinctively, evolutionary algorithms through global search can discover exciting rules and patterns that would be missed by the greedy search performed by most rule induction methods [3].

2. Related work:

Ee-Peng Lim and Aixin Sun [5] represent ontology as a set of concept and their inter-relationships applicable to some information domain. The knowledge provide by ontology is enormously helpful in defining the formation and possibility for Web content mining. They also made review of Web mining and describe the ontology techniques to Web mining. The relevance of these Web mining techniques to digital library systems was also taken care of.

B. Rajdeepa, DR. P. Sumathi, [6] (2014), projected the Web Structure according to an effectiveness criterion. The frequency procedure and time separation to tick on a link were

used for constructing the web user utility. Discussed that doing research on behavioural probability has allowed quality growth of the hyperlink structure, using a Hybrid GA/PSO. The discussed methodology was tried on a real web site and the results were to be interpreted using the web user benefits by links. The simplicity and effectiveness of Particle Swarm Optimization technique based on the idea of Swarm cleverness was executed in high-dimensional order clustering investigation for web usage mining. As the Hybrid GA/PSO algorithm has fast convergence and simple achievement, numerous enhanced versions of Hybrid GA/PSO algorithm have been developed to resolve that difficulty of GA and PSO.

Soumadip Ghosh, Amitava Nag, Debasish Biswas, Arindrajit Pal, [7] (2011), have dealt with a challenging association rule mining problem of finding frequent itemsets from XML database using proposed GA based method. The method, described here is very simple and efficient. It was successfully tested for different XML databases. The results reported are correct and appropriate. Also tried to compare the performance of GA based method proposed with the FP-tree algorithm.

V.V.R. Maheswara Rao, Dr. V. Valli Kumari and Dr. K.V.S.V.N Raju, [8] (2010), introduced the Advanced Optimal Web Intelligent Model with granular computing nature of Genetic Algorithms, IOG. The IOG model is designed on the semantic enhanced content data, which works more efficiently than that on normal data. The unique parameters of the IOG fitness function generates balanced weights, to represent the significance of characteristics, which yields the dissimilar characteristics of page vector. The evaluation function of IOG improves the page quality and reduces the execution time to a specified number of iterations as it considers practical measures like page, link and mean qualities. The genetic operators are intended in drawing the stickiness among the characteristics of a page vector. By integrating all the above the web intelligent model IOG, it significantly improves the investigation of web user usage behaviour.

Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha Pratim Sarkar, [9] (2010), dealt with a challenging association rule mining problem of finding frequent itemsets using proposed GA based method. The method, described is very simple and efficient. This is successfully tested for different large data sets. The results reported are correct and appropriate. However, a more extensive empirical evaluation of the proposed method will be the objective of future research. Compared the performance of GA based method proposed with the FP-tree algorithm.

3. Methodology:

Genetic Algorithms (GA) are direct, parallel, stochastic method for global search and optimization, which imitates the evolution of the living beings, described by Charles Darwin. GA are part of the group of Evolutionary Algorithms (EA). The evolutionary algorithms use the three main principles of the natural evolution: reproduction, natural selection and diversity of the species, maintained by the differences of each generation with the previous. Genetic Algorithms works with a set of individuals, representing possible solutions of the task. The selection principle is applied by using a criterion, giving an evaluation for the individual with respect to the desired solution. The best-suited individuals create the next generation. The large variety of problems in the engineering sphere, as well as in other fields, requires the usage of algorithms from different type, with different characteristics and settings.

Genetic algorithms are a type of optimization algorithm, meaning they are used to find the optimal solution to a given computational problem that maximizes or minimizes a particular function. Genetic algorithms represent one branch of the field of study called evolutionary computation, in that they imitate the biological processes of reproduction and natural selection to solve for the `fittest' solutions [1]. Like in evolution, many of a genetic algorithm's processes are random, however this optimization technique allows one to set the level of randomization and the level of control [1]. These algorithms are far more powerful and effcient than random search and exhaustive search algorithms, yet require no extra information about the given problem. This feature allows them to find solutions to problems that other optimization methods cannot handle due to a lack of continuity, derivatives, linearity, or other features.

A. How Genetic Algorithm Works

The genetic algorithm is a technique to solve constrained and unconstrained optimization problems both, based on natural selection, a system that caters to biological evolution. The genetic algorithm constantly changes population of individual solutions. At each stage, the genetic algorithm selects individuals at random from the current population to be parents and uses them to produce the children for the next generation. Over succeeding generation, the population "evolves" toward an most favourable solution. The genetic algorithm can be used to solve diversity of optimization problems that are not suited to standard optimization algorithms, inclusive of problems where in the objective functions are discontinuous, non-differentiable, stochastic, or highly nonlinear.

The genetic algorithm optimization techniques makes use of three important rules for the creation of the next generation from the current population. These are

Selection rules : It selects the individuals, called parents, that donate to the population at the next generation.

Crossover rules: It combines two parents to form children for the next generation.

Mutation rules: It applies random changes to individual parents to form children.

Given below is the brief description as to how genetic algorithm wortks:

It begins by creating a random initial population. Next it creates a series of new populations.
Netx, at each and every step the algoriothm uses the individuals in the present generation to develop the next population.
For creation of new population, the GA performs the following process:

1. The member of the current population is given a score by calculating its fitness value.
2. Then the scorres so obtained are scaled to a more appropriate range.
3. Next, to identify the members, called parents, based on the current fitness values.
4. Parents in the present population having less fitness values are called as elite, who are passed on to the next population.
5. Next, children are produced from the parents, either by making arbitrary changes to a single parent—mutation—or by combining the vector entries of a pair of parents—crossover.
6. In order to form the next generation, the current population is replaced by the children.
7. Finally to stop the algorithm when the stopping criteria is reached.

4 Model Inputs and Algorithm Development

The modelling approach used to develop forecasting model along with details on input and output parameters is presented in this section. One of the most important steps in the development of any prediction model is the selection of appropriate input variables that will allow GA to successfully produce the desired results. Good understanding of the system under consideration is an important prerequisite for successful application of data driven approaches. The main reason for this is that GA belongs to the class of data-driven approaches. Physical understanding of the process being studied leads to better choice of the input variables. Here since predicting web usage behaviour is a complicated problem that involves multiple interacting factors. In order to build a reasonably accurate model for prediction, proper parameters must be selected. Some practical considerations in parameter selections are firstly, the selected parameters must affect the target problem, i.e., strong relationships must exist among the parameters and target (or output) variables, and secondly, the selected parameters must be well-populated, and corresponding data must be as clean as possible. Since the soft computing methods model problems based on available data, the availability and quality of data are both essential.

A. Dataset Used

In the present work the data that is going to be used is the Online News Popularity dataset available from UCI Machine Learning Repository [1] made publicly available for carrying out further research work in the field of software engineering for predictive model development. The original source of the dataset is "K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News" in the Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal. It gives a basic overview of the dataset. It has 58 predictive attributes, 2 non-predictive, 1 goal field. Further, in the present thesis for model development out of 58 attributes, only 3 have been used for model development. The dataset features are given in Table-1 below. Some example of input given in data is like the number of images (num_imgs) and the number of links (num_hrefs). The output of the model is the number of shares (shares**)** of the news sites. Data are going to be used for predicting the popularity of the news channels using the number of shares of different online news channels using GA. For developing the GA models MATLAB platform will be used.

Table 1: List of some Attributes given for model development

Sl. No.	Attribute Name	Input/Output Variable
1	***num_hrefs: Number of links***	As input
2	**num_imgs: Number of images**	As input
3	**n_non_stop_words**	As input
4	**num_keywords**	As input
5	***shares: Number of shares***	As Output

After optimization of the fitness function using MATLAB command, the optimized function value and the optimal parameter values are obtained. Using different parameter options for GA algorithm functions solutions obtained are as follows:

$S_{estimated}$ = a+ *b*x(1) + c*x(2);* *(3)*

Where , a=10.0476472926427 ,b= 12.8924554769075 and c= 23.0108234662576

Further the figure 1 below shows optimized values of all the datasets actual and estimated output using GA optimization.

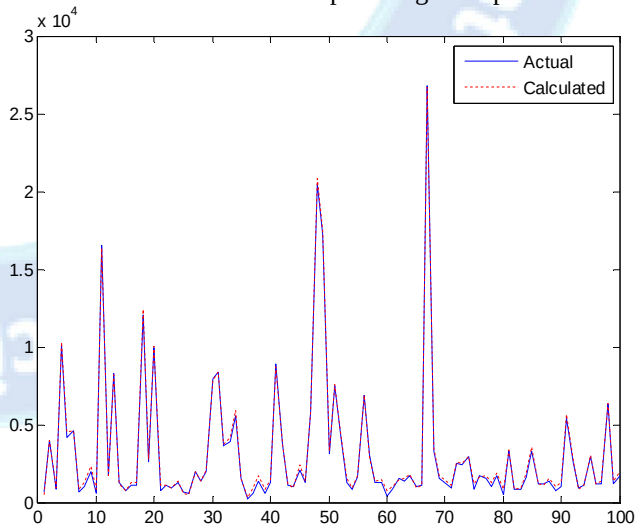

Fig. 1. Plot of estimated results for all the datasets

On further analysis of the above equations (2) and (3), it was seen that equation (3) for the model was found to be the best developed model, resulting in low RMSE value of 12.90 or less as compared to that of regression model for the same datasets having RMSE value. It clearly demonstrates that genetic algorithm optimization techniques has been successful in developing a better prediction model by lowering the RMSE value. It is shown in figure 1. The Matlab plot of the various dataset results in terms of actual and calculated are generated in the optimization of the model has been shown in figure 1 and as magnified view in 1.The the confidence value for above mentioned a,b and c parameter values is 71.33%.

5. Conclusion:

In this study, applicability and capability of Genetic Algorithm techniques for application in web mining as a predictive tool has been investigated. It is seen that GA models are very robust, characterised by fast computation, capable of handling the noisy and approximate data that are typical of data used here for the present study. From the analysis of the results given earlier it is seen that GA has been able to perform well for the prediction of the number of shares of the various news popularity sites.

References:

[1] M. Berry and G. Lino_, Data Mining Techniques. For Marketing, Sales and Customer Support, John Wiley & Sons, Inc. 1997.

[2] R. Agrawal, T. Imielinski and A. Swami, \Mining Association Rules between Sets of Items in Large Databases", In Proc. of the 1993 ACM SIGMOD International Conference on Management of Data, 1993.

[3] A. Freitas, \Data Mining and Knowledge Discovery with Evolutionary Algorithms", Springer-Verlag, New York Inc., 2002.

[4] A. A. Freitas, \Data Mining and Knowledge discovery with Evolutionary Algorithms", IEEE Transactions on Evolutionary Computation, Vol. 7, No. 6, pp. 517-518, 2003.

[5]. Ee-Peng Lim and Aixin Sun, WEB MINING - THE ONTOLOGY APPROACH, pp. 1-8.

[6]. B. RAJDEEPA, DR. P. SUMATHI, (2014), "WEB STRUCTURE MINING FOR USERS BASED ON A HYBRID GA/PSO APPROACH", Journal of Theoretical and Applied Information Technology, Vol.70 No.3, pp. 573-578.

[7] Soumadip Ghosh, Amitava Nag, Debasish Biswas, Arindrajit Pal, (2011), " XML MINING USING GENETIC ALGORITHM", Journal of Global Research in Computer Science, Volume 2, No. 5, pp. 86-90.

[8] V.V.R. Maheswara Rao, Dr. V. Valli Kumari and Dr. K.V.S.V.N Raju, (2010), "An Advanced Optimal Web Intelligent Model for Mining Web User Usage Behavior using Genetic Algorithm", ACEEE Int. J. On Network Security, Vol. 01, No. 03, Dec 2010, pp. 44-49.

[9] C.J. Carmona, S. Ramírez-Gallego, F. Torres, E. Bernal, M.J. del Jesus, S. García, (2012), "Web usage mining to improve the design of an e-commerce website: orolivesur.com", Expert Systems with Applications 39, 11243–11249.